Quantifying the Relationship between Visual Salience and Visual Importance

Junle Wang[1,2], Damon M. Chandler[3], and Patrick Le Callet[2]

[1] School of Electronic and Information Engineering,
South China University of Technology, Guangzhou, China

[2] IRCCyN UMR no 6597 CNRS,
Ecole Polytechnique de l'Universite de Nantes, Nantes, France

[3] School of Electrical and Computer Engineering,
Oklahoma State University, Stillwater, OK USA

ABSTRACT

This paper presents the results of two psychophysical experiments and an associated computational analysis designed to quantify the relationship between visual salience and visual importance. In the first experiment, importance maps were collected by asking human subjects to rate the relative visual importance of each object within a database of hand-segmented images. In the second experiment, experimental saliency maps were computed from visual gaze patterns measured for these same images by using an eye-tracker and task-free viewing. By comparing the importance maps with the saliency maps, we found that the maps are related, but perhaps less than one might expect. When coupled with the segmentation information, the saliency maps were shown to be effective at predicting the main subjects. However, the saliency maps were less effective at predicting the objects of secondary importance and the unimportant objects. We also found that the vast majority of early gaze position samples (0-2000 ms) were made on the main subjects, suggesting that a possible strategy of early visual coding might be to quickly locate the main subject(s) in the scene.

1. INTRODUCTION

Visual salience[1, 2] and visual importance[3–6] can provide important insights into how biological visual systems address the image-analysis problem. However, despite the differences in the way (bottom-up) visual salience and (top-down) visual importance are determined in terms of human visual processing, both salience and importance have traditionally been considered synonymous in the signal-processing community: They are both believed to denote the most visually "relevant" parts of the scene. In this study, we present the results of two psychophysical experiments and an associated computational analysis designed to quantify the relationship between visual salience and visual importance.

A psychophysical experiment was performed to obtain visual importance maps for a large database of images. A visual importance map is an object-level map that specifies the visual importance of each object in an image relative to the other objects in the image (including what would normally be considered the background). The object(s) that receive the greatest visual importance are traditionally considered the image's main subject(s). By using images from the Berkeley Image Segmentation Dataset, we collected importance ratings for each object in the 300 database images. Such importance ratings are generally believed to result from top-down visual processing since the decisions used to rate each object typically involve scene interpretation, object recognition, and oftentimes consideration of artistic intent.

In a second experiment, visual gaze patterns were measured for 80 of the images from the same Berkeley Image Segmentation Dataset. Using an eye-tracker, visual gaze locations were recorded under task-free viewing. Whereas importance maps are driven primarily by top-down processing, visual gaze patterns are generally believed to be driven by bottom-up, signal-based attributes, at least for early gaze locations. Bottom-up saliency[1] is one particular signal-based attribute which has been shown to correlate well with early gaze locations. An image region is considered visually salient if it "stands out" from its background in terms of one or more attributes (e.g., contrast, color). When visual gaze patterns are measured in task-free viewing, one can consider the locations

J.W.: E-mail: junle.wang@etu.univ-nantes.fr; D.M.C.: E-mail: damon.chandler@okstate.edu; P.L.: E-mail: patrick.le-callet@univ-nantes.fr

to denote the salient regions in the image. Thus, from the gaze patterns, one can construct an experimental saliency map.

The use of the same images in both experiments allows us to perform a computational analysis of the saliency maps and importance maps to quantify their similarities and differences. A similar analysis was earlier performed by Engelke *et al.* in which visual fixation patterns were compared to human-selected regions of interest (ROIs).[7] Engelke *et al.* concluded that there indeed exists a relationship between visual fixation patterns and ROIs, with early fixations demonstrating a stronger relationship than later fixations. Here, we perform a related study using importance maps rather than ROIs, with a particular emphasis on quantifying any potential relationships as a function of time.

This paper is organized as follows: Section 2 describes the experimental methods used to collect the saliency maps and importance maps. Section 3 presents the results and analyses of the experiments. A discussion of our findings is provided in Section 4. General conclusions are presented in Section 5.

2. METHODS

Two psychophysical experiments were performed to obtain saliency maps and importance maps. In Experiment I, subjective ratings of importance were recorded to obtain importance maps. In Experiment II, visual gaze patterns were recorded to obtain saliency maps.

2.1. Experiment I: Visual Importance

In Experiment I, subjective ratings of importance were obtained for each object in each of 300 images to obtain importance maps. The methods were as follows.

Apparatus: Stimuli were displayed on a ViewSonic VA912B 19-inch LCD monitor (1280×1024 at 60 Hz). The display yielded minimum and maximum luminance of respectively, 2.7 and 207 cd/m^2. Stimuli were viewed binocularly through natural pupils in a darkened room at a distance of approximately 1575 pixels through natural pupils.

Stimuli: Stimuli used in Experiment I were obtained from the Berkeley Segmentation Dataset and Benchmark image database. This database was chosen because its images are accompanied by human-segmented versions (averaged over at least five subjects). We hand-segmented all 300 images in the database into 1143 objects by using the database's hand-segmented results as a reference (the database provides only edge-map segmentations rather than object-map segmentations). The images used were 321×481 and 481×321 pixels with 24-bit RGB values.

Procedures: For each of the 1143 objects, subjects were instructed to rate the perceived importance relative to the other objects within the image. The ratings were performed using an integer scale of 0 to 10 in which 10 corresponded to greatest importance and 0 corresponded to least importance. The time-course of each session was not limited; however the majority of subjects completed the experiment in less than 120 minutes.

Raw scores for each subject were converted to z-scores. The per-subject z-scores were then averaged across all subjects, and then the average z-scores were rescaled to span the range $[0, 1]$ for each image. From these results, a per-image *importance map* was obtained by assigning each object's average importance score to all pixels in that object. Thus, in each importance map, brighter regions denote objects of greater visual importance.

Subjects: Ten adult subjects participated in the experiment. Three of the subjects were familiar with the purpose of the experiment; the other subjects were naive. Subjects ranged in age from 21 to 34 years. All had either normal or corrected-to-normal visual acuity.

2.2. Experiment II: Visual Salience

In Experiment II, an eye-tracker was employed to measure visual gaze points and thereby compute visual saliency maps using 80 of the images used in Experiment I. The methods were as follows.

Apparatus: Stimuli were displayed on a Dell 1704FPT LCD 17-inch monitor (1280×1024 at 60 Hz). The display yielded minimum and maximum luminance of respectively, 0.3 and 180 cd/m^2. Stimuli were viewed binocularly through natural pupils in a darkened room at a distance of approximately 1575 pixels through

natural pupils. Eye-tracking was performed by using the Video Eyetracker Toolbox from Cambridge Research Systems. This system tracks via acquisition/processing of pupil and dual first Purkinje images. The system has an accuracy of 0.25° and a sampling rate of 50 Hz.

Stimuli: The stimuli used in Experiment II consisted of 80 of the 300 images used in the Experiment I. The 80 images were selected to contain only the landscape orientation (481×321), and to provide a uniform sampling in terms of the number of objects per image and the levels of importance per image.

Procedures: A task-free viewing paradigm was employed in which subjects were instructed simply to look at the images given no specific task. Each of the 80 stimuli was presented for 15 seconds. The order of the presentation was randomized for each observer. Calibration of the eye-tracker was performed periodically throughout the approximately 30-minute experimental session.

From the gaze position samples, we constructed a per-image *saliency map* by placing a two-dimensional Gaussian at each gaze sample point from all observers. The standard deviation of the Gaussian was determined based to the size of fovea. We then normalized all saliency maps to span the range $[0, 1]$.

Subjects: Eighteen adult subjects participated in the experiment. All subjects were paid participants and were naive to the purpose of the experiments. Subjects ranged in age from 19 to 45 years. All had either normal or corrected-to-normal visual acuity.

3. RESULTS AND ANALYSIS

3.1. Qualitative Observations of Importance Maps and Saliency Maps

A qualitative comparison of the saliency maps and importance maps reveals some distinct similarities and differences between the two. Figure 1 depicts some representative examples.

The importance maps suggest that object category plays a bigger role than most other factors in determining subjective importance. In general, we found that subjects tended to rate objects containing human faces and/or animals to be of greatest importance. Background objects such as sky and grass were generally rated to be of least importance. Occlusion (i.e., whether an object is in the foreground vs. the background) also seems to be an important factor for perceived importance.

The saliency maps generally suggest that regions which possess a distinguished shape, color, contrast, or other local spatial features attract attention. However, subjects always gazed upon the image's main subject(s): Gaze position samples tended to occur on objects which belong to animal faces, human faces, or other subjects which represent the gist of the image. The background, such as sky and ground, always attracted the least attention.

Yet, despite these similarities, the saliency maps and importance maps don't always agree. Although we employed a relatively long viewing period, the saliency maps never yielded an object-level segregation which is enforced in the importance maps. For example, whenever a face occurred in an image, whether an animal face or a human face, the subjects' gaze positions always occurred on the face. Furthermore, as demonstrated by the bottom-most image the left-hand group in Figure 1, which contains people in the background (the people are located toward the top-left corner of the image), the importance ratings seem to be influenced by the artistic intent (e.g., the flowers and reflecting pond), whereas the saliency maps highlight only some of these regions.

3.2. Predicting the Main Subject, Secondary Objects, and the Background

The results of the qualitative analysis suggest a relationship between saliency maps and importance maps. One way to quantify this relationship is to attempt to predict the importance maps from the saliency maps using the object-level segmentations as side-information. This approach decouples the errors in prediction from the errors due to segmentation, since the latter would otherwise give rise to an artificially low measure of dependency between the two maps.

To predict the importance maps from the saliency maps (given the segmentations), the following two approaches were tested:

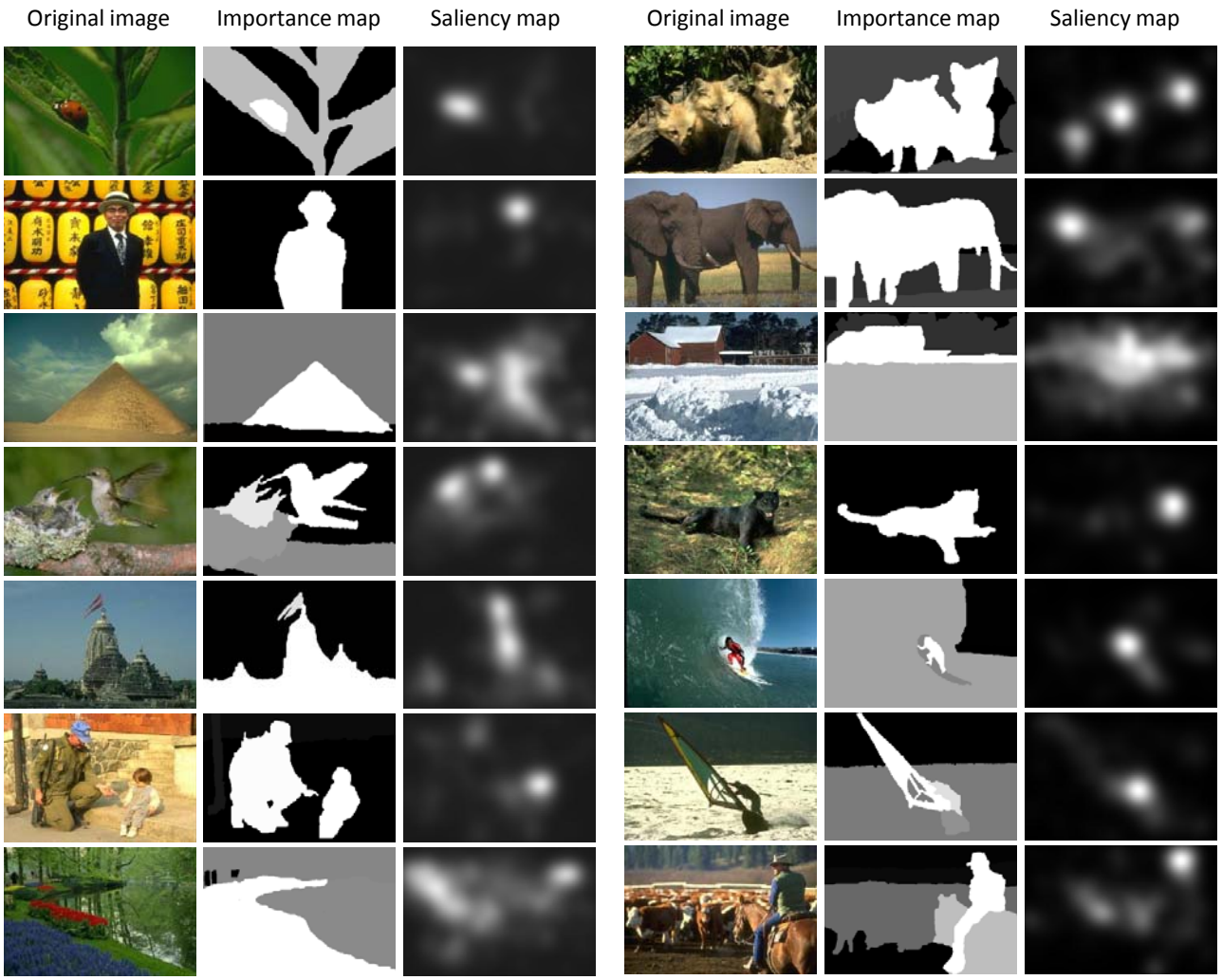

Figure 1. Representative results from the experiments.

1. *Mean Saliency*: For each object, we summed those values of the saliency map which occurred within the object, and then we divided this value by the total number of pixels in the object. For each image, the resulting set of per-object saliency values was then normalized to span the range $[0, 1]$.

2. *Coverage Saliency*: For each object, we summed those values of the saliency map which occurred within the object, and then we divided this value by the number of pixels in the object that were gazed upon (specifically, the number of pixels that were covered by the fovea). For each image, the resulting set of per-object coverage saliency values was then normalized to span the range $[0, 1]$.

To facilitate the prediction, each importance map was quantized into three classes based on the importance values: (1) *Main subjects*, which consisted of objects which received an importance value ranging from 2/3 to 1; (2) *Secondary objects*, which received an importance value ranging from 1/3 to 2/3; (3) *Background objects*, which received an importance value ranging from 0 to 1/3.

The results of the prediction are provided in Table 1 in the form of confusion matrices. Each row of each matrix represents the actual importance class, and each column represents the predicted class. An ideal prediction would yield a diagonal matrix with 100% values. As shown in Table 1(a), the average saliency can successfully predict the main subject approximately 81% of the time. Similarly, the background is successfully predicted

	Main Sbj.	Secondary	Background
Main Sbj.	**80.5%**	29.8%	12.6%
Secondary	12.5%	**42.6%**	40.7%
Background	7.1%	27.6%	**46.7%**

(a) Using Mean Saliency

	Main Sbj.	Secondary	Background
Main Sbj.	**56.5%**	38.6%	8.2%
Secondary	13.0%	**40.4%**	24.7%
Background	30.5%	21.1%	**67.1%**

(b) Using Coverage Saliency

Table 1. Confusion matrices for predicting each object's importance from the gaze data.

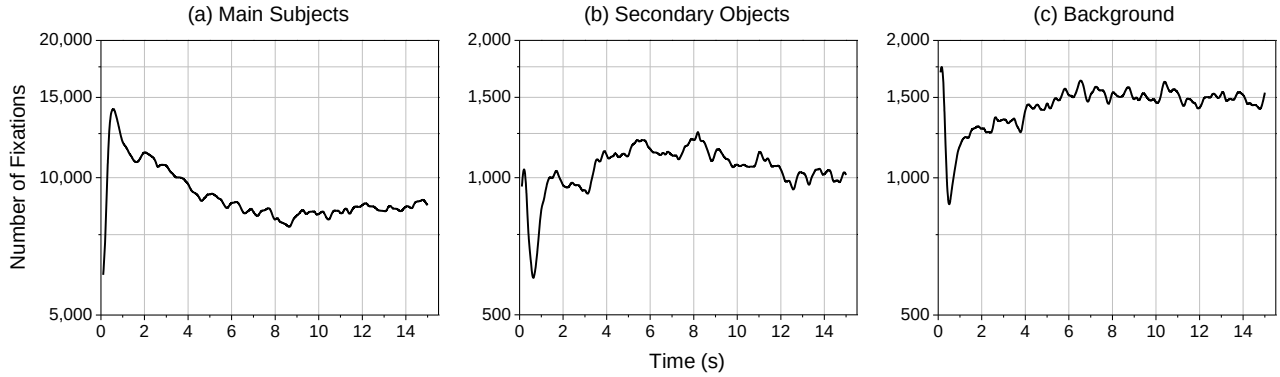

Figure 2. Total number of gaze position samples in (a) main subjects, (b) secondary objects, and (c) background objects computed in each 100-ms interval of the 15-second viewing time. Note that the scale for the vertical axis in the first graph is 10x that of the other two graphs.

approximately 47% of the time.* Coverage saliency, shown in Table 1(b), yields worse performance for main subjects, but slightly better performance for background objects.

3.3. Temporal Analysis

During normal viewing, because visual attention shifts from one object to another, the number of gaze position samples which occur on each object varies according to time. For each of the three levels of importance (main subjects, secondary objects, background), we analyzed this time dependence. Specifically, we computed the number of gaze position samples per importance class which occurred within each 100-ms interval during the 15-second viewing time. The resulting three time curves, summed across all observers, are shown in Figure 2.

The plots in Figure 2 clearly indicate that, on average, objects from different importance classes attract considerably different amounts of visual attention. Specifically, throughout the 15-second viewing time, the main subjects always received the greatest number of gaze position samples—approximately 7-8 times greater than the number of samples for secondary and background objects.

Within 0-500 ms, the number of gaze position samples for the main subjects was already 4-6 times greater than the number of samples for secondary and background objects. This observation suggests bottom-up mechanisms can be effective at locating the main subjects in these images; this might result from the fact that photographers tend to increase the saliency of the main subjects via, e.g., retouching, selective focusing, or other photographic

*Note, however, that some objects contain only two levels of importance: main subject and background.

	Main Sbj.	Secondary	Background
Main Sbj.	**89.0%**	43.5%	12.4%
Secondary	3.3%	**43.5%**	27.2%
Background	7.7%	13.0%	**60.5%**

Table 2. Confusion matrix for predicting importance from the first 2 sec. of gaze samples using mean saliency.

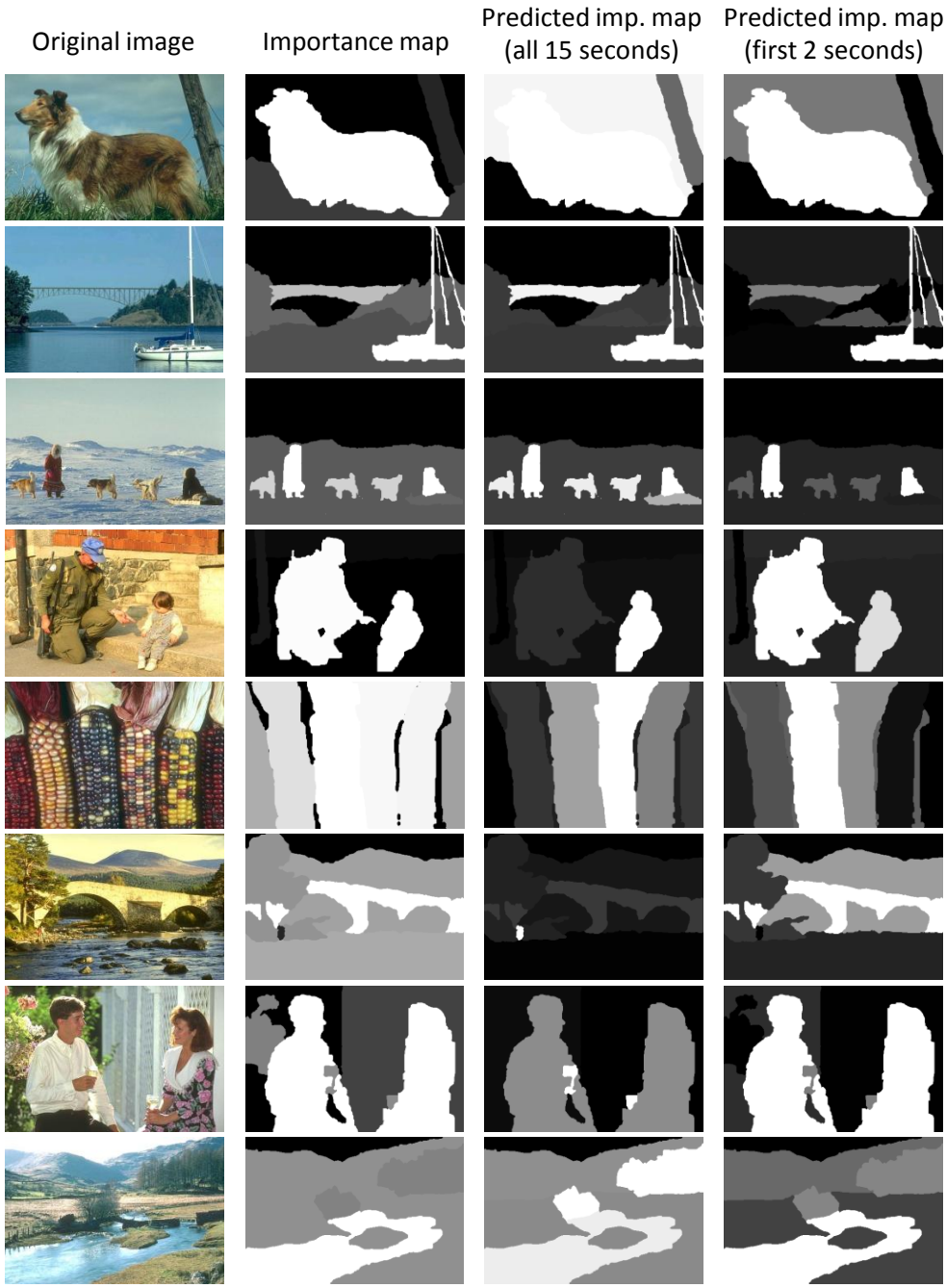

Figure 3. Representative results of using all gaze samples vs. only those from the first two seconds to predict the importance maps.

techniques. Between 500-2000 ms, there was a pronounced increased in the number of gaze position samples for the main subjects, while the number for the other two importance classes decreased in this period. These changes potentially indicate the influence of top-down mechanisms which might force observers to attend to the main subjects. After this process, the number of gaze position samples for the main subjects slightly decreased and those for the other two classes slightly increased. This latter change may imply that the observers attempt to explore the whole image, but their attention is still held by the main subjects.

These three time curves suggest that the relationship between visual salience and visual importance may

Figure 4. Examples of subjects visually attending to only select regions of objects, most notable for images containing faces.

be time dependent. In particular, the fact that the main subjects attract the most attention within 0-2000 ms suggests that these early gaze position samples might be a better predictor of visual importance for the main subjects than previously achieved using all samples. Accordingly, we predicted the importance maps by using the samples taken from only the first 0-2000 ms. Table 2 lists the resulting confusion matrix computed using mean saliency as the prediction mechanism from the first two seconds of gaze data. Figure 3 depict representative importance maps predicted from the gaze data taken from all 15 seconds and from only the first two seconds. By using only these early gaze data, better prediction is achieved for the main subjects.

4. DISCUSSION

In this study, we have attempted to quantify the similarities and differences between bottom-up visual salience and top-down visual importance. The implications of these initial findings for image processing are quite important. Several algorithms have been published which can successfully predict gaze patterns (e.g., Refs. 2, 8). Our results suggest that these predicted patterns can be used to predict importance maps when coupled with a segmentation scheme. In turn, the importance maps can then be used to perform importance-based processing such as auto-cropping, enhancement, compression, unequal error protection, and quality assessment.

Yet, as we have seen here, the predictions are not as great as one might have expected. Below we describe some possible explanations along with some implications for visual coding.

4.1. The Segmentation Problem

When humans look at an interesting object under task-free viewing, they do not fixate upon all the areas of the object. Only select regions of an object attract the vast majority of visual fixations. The clearest examples of these are for human/animal faces vs. the rest of their bodies; for most images, people looked only at the faces. Figure 4 depicts some prime examples of this observation.

Despite the fact that only select regions of an object attract the vast majority of visual fixations, these objects were rated by subjects to be of high importance. This suggests that either: (1) humans tend to base the importance of the entire object on the importance of the most interesting region (max operator under visual summation); or (2) some objects could benefit from another level of segmentation. In Experiment I, subjects were forced to rate object-level importance, where the object was defined by the given segmentation. One could possibly obtain ratings for each object's sub-objects and draw a different conclusion. Although the segmentations used in Experiment I were based on hand-segmentations provided by the Berkeley dataset, the results of our analyses are restricted by these segmentations.

4.2. The Border-Ownership Problem

When a gaze samples falls on or near the border of an object, there remains a question as to which object the gaze sample should be assigned. Indeed, perception itself dictates which object owns the border (the "border ownership" problem; Ref. 9).

Figure 5. Examples of ambiguities regarding to which object a gaze position sample should be assigned.

For example, as shown in Figure 5(a), the gaze position samples tend to suggest that the subjects were visually attending to the whale, despite the fact that some of the near-border gaze samples actually fall upon the background (water). However, as shown in Figure 5(b), it is not always easy to decipher to which object the subjects were visually attending. In this latter case, it is not clear whether near-border samples should be assigned to the men or the elephants.

Because of this ambiguity, the importance maps for images such as that depicted in Figure 5(a) may reflect abnormally high values of importance for the background. In terms of border ownership, the region to which the border is assigned is perceived as the figure. However, without knowledge of each subject's instantaneous figure/ground interpretation, there is no correct way to assign a gaze position sample to a particular object.

4.3. Bottom-Up vs. Top-Down Visual Coding

Two mechanisms are at work when humans look at an image: bottom-up and top-down. The results of our temporal analyses suggest that visual attention varies according to time and that the extent to which visual attention and visual importance maps are related also varies with time.

The results of the temporal analysis revealed that the relationship between visual attention and visual importance is strongest in the first two seconds of the 15-second observation interval, which implies that top-down mechanisms dominate eye movements during this period. This finding was also confirmed by attempting to predict the importance class from the saliency map generated from the early gaze samples. A good prediction for the primary ROI implies that top-down mechanisms dominate eye movements (observers are looking at the most important objects), whereas a bad prediction suggests either that bottom-up mechanisms are at work, or that the two mechanisms are competing.

These results suggest a possible strategy for human visual coding. If the human visual system can so rapidly identify the main subject(s) in a scene, such information can be used to prime lower-level neurons to better encode the visual input. This strategy may play an important role toward improving neural coding efficiency or facilitating higher-level tasks such as scene categorization. Several researchers have suggested that rapid visual priming might be achieved via feedback and/or lateral interactions between groups of neurons after the "gist" of the scene is determined.[10,11] The results of this study provide psychophysical evidence that lends support to a gist-based strategy and a possible role for the feedback connections that are so prevalent in mammalian visual systems.

5. CONCLUSIONS

This paper presented the results of two psychophysical experiments and an associated computational analysis designed to quantify the relationship between visual salience and visual importance. We found that saliency maps and importance maps are related, but perhaps less than one might expect. The saliency maps were shown to be effective at predicting the main subjects. However, the saliency maps were less effective at predicting the objects of secondary importance and the unimportant objects. We also found that the vast majority of early gaze position samples (0-2000 ms) were made on the main subjects. This suggests that a possible strategy of the human visual system is to quickly locate the main subject(s) in the scene.

There are several ways in which one may be able to improve the prediction. The timing data suggests that a multi-stage predictor might be more effective than our single-stage attempts. In the first stage, the gaze position samples from 0-2000 ms (or another early time slice) would be used to determine the most important objects (main subjects). This knowledge of the main subjects would then be used to guide subsequent predictions for other objects. We are currently implementing this scheme.

REFERENCES

1. C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry.," *Hum Neurobiol* **4**(4), pp. 219–227, 1985.
2. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), pp. 1254–1259, 1998.
3. W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image," pp. 701–704, 1998.
4. A. Maeder, "Importance maps for adaptive information reduction in visual scenes," in *Intelligent Information Systems, 1995. ANZIIS-95*, pp. 24–29, 1995.
5. S. P. Etz and J. Luo, "Ground truth for training and evaluation of automatic main subject detection," *Human Vision and Electronic Imaging V* **3959**(1), pp. 434–442, SPIE, 2000.
6. V. Kadiyala, S. Pinneli, E. C. Larson, and D. M. Chandler, "Quantifying the perceived interest of objects in images: effects of size, location, blur, and contrast," in *Human Vision and Electronic Imaging XIII*, T. N. Rogowitz, Bernice E.; Pappas, ed., **6806**, pp. 68060S–68060S–13, 2008.
7. U. Engelke, H. J. Zepernick, and A. Maeder, "Visual attention modeling: Region-of-interest versus fixation patterns," in *Proc. 27th Picture Coding Symposium*, 2009.
8. O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(5), pp. 802–817, 2006.
9. K. Nakayama, S. Shimojo, and G. H. Silverman, "Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects," *Perception* **18**, pp. 55–68, 1989.
10. A. Friedman, "Framing pictures: The role of knowledge in automatized encoding and memory for gist," *Journal of Experimental Psychology: General* **108**, pp. 316–355, 1979.
11. A. Oliva, "Gist of the scene," in *The Encyclopedia of Neurobiology of Attention*, L. Itti, G. Rees, and J. K. Tsotsos, eds., pp. 251–256, Elsevier, San Diego, CA, 2005.